ЗШ

УДК 004.934.1’1

В.Ю. Шелепов, А.В. Ниценко, Г.В. Дорохина
Институт информатики и искусственного интеллекта
ГВУЗ «Донецкий национальный технический университет», г. Донецк, Украина
Украина, 83050, г. Донецк, пр. Б. Хмельницкого, 84
Институт проблем искусственного интеллекта МОН Украины и НАН Украины
Украина, 83048, г. Донецк, ул. Артема, 118-б

О распознавании речи на основе межфонемных переходов

V.Ju. Shelepov, A.V. Nicenko, G.V. Dorohina
Institute of *Informatics and* Artificial Intelligence
of *Donetsk National Technical University, Donetsk, Ukraine*
Ukraine, 83050, c. Donetsk, B. Khmelnitskiy st., 84
Institute of Artificial Intelligence MES of Ukraine and MAS of Ukraine
Ukraine, 83048, c. Donetsk, Artema st., 118-b

On Speech Recognition Using Phoneme Transition Base

В.Ю. Шелєпов, А.В. Ніценко, Г.В. Дорохіна
Інститут інформатики і штучного інтелекту
ДВНЗ «Донецький національний технічний університет», м. Донецьк, Україна
Україна, 83050, м. Донецьк, пр. Б. Хмельницького, 84
Інститут проблем штучного інтелекту МОН України і НАН України
Україна, 83048, м. Донецьк, вул. Артема, 118-б

Про розпізнавання мовлення на підставі межфонемных переходів

Содержание статьи относится к области пофонемного распознавания русских слов. В качестве базовых элементов предлагается использовать не стационарные части звуков речи, а участки межфонемных переходов. Описываются преимущества такого подхода. Предлагается метод распознавания слов по эталонам, автоматически формируемым программой с использованием заранее созданной базы межфонемных переходов. Важнейшее применение – распознавание больших словарей. Описывается программа быстрого создания базы межфонемных переходов на основе автоматической сегментации речевого сигнала.
Ключевые слова: межфонемный переход, дифон, сегментация,
распознавание больших словарей

The content of the article relates to phoneme recognition of Russian words. The aim is to propose the method for word recognition by samples, which program automatically creates phoneme transition base. The most important application is large vocabularies recognition.
Key words: phoneme transition, diphone, segmentation, large vocabularies recognition

Зміст статті належить до сфери пофонемного розпізнавання російських слів. Як базові елементи пропонується використовувати не стаціонарні частини звуків мовлення, а відрізки міжфонемних переходів. Наводяться переваги такого підходу. Пропонується метод розпізнавання слів за еталонами, які автоматично формуються програмою з використанням заздалегідь створеної бази міжфонемних переходів. Найважливіше застосування – розпізнавання великих словників. Описується програма швидкого створення бази міжфонемних переходів на основі автоматичної сегментації мовленнєвого сигналу.
Ключові слова: міжфонемний перехід, дифон, сегментація,
розпізнавання великих словників

I. Стремясь создать систему пофонемного распознавания русских слов, авторы данной работы долгое время пытались использовать в качестве элементов распознавания стационарные части звуков речи [1-2]. К этому нас побуждало то, что общее количество таких звуков (гласных, звонких согласных, шипящих, аффрикат и т.д.) не велико, всего несколько десятков. Однако хорошо известен эффект коартикуляции – влияние друг на друга соседних звуков. Например, согласный звук заметно меняется, если за ним следует огубленный гласный [о] или [у]. Учитывая важнейшую роль пар соседних звуков в слове, авторы решились бы сформулировать свою сегодняшнюю точку зрения в виде следующего тезиса: *Один из возможных ключей к распознаванию речи лежит в межфонемных переходах.*

II. Анализ высказанного утверждения можно начать со следующего простого эксперимента. Используя какую-либо известную программу работы со звуком, например «Sound Forge», запишем два произвольных слова, а затем вырежем стационарные (серединные) части составляющих их звуков. Воспроизведя получившийся звуковые сигналы, мы можем на слух определить, какие слова звучат. Напротив, вырезав межфонемные переходы и оставив стационарные части фонем, мы затруднимся на слух различить, например, слова «мама» и «лама».

Следующий аргумент относительно роли при распознавании межфонемных переходов – использование при DTW-распознавании [3], [2] эталонов слов, полученных удалением стационарных частей звуков, из которых эти слова состоят. Эксперименты показывают, что такое распознавание не менее успешно, чем распознавание по «полным эталонам». На рис. 1 приведено окно для DTW- распознавания.

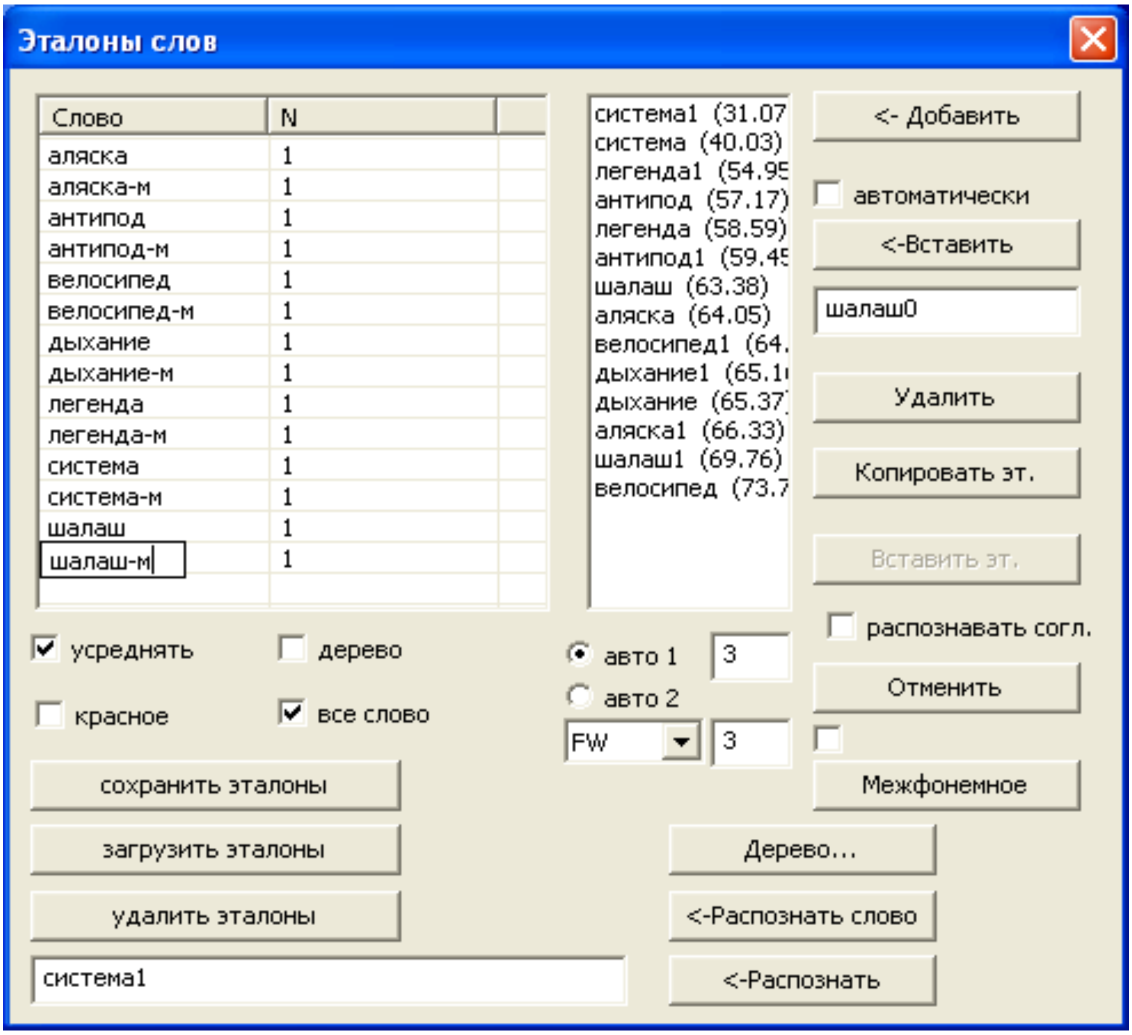


Рисунок 1 – Окно для DTW-распознавания с полными и урезанными эталонами

В левом списке расположены слова, для которых построены эталоны (о них чуть ниже) по полному сигналу, и те же слова (они снабжены в конце символом -м), которые получены оставлением лишь межфонемных переходов – 3 окна по 368 отсчетов слева и 3 окна по 368 отсчетов справа от метки между соседними звуками (это достигается нажатием кнопки «межфонемное». Количество окон указывается в полях справа. После этого эталоны строятся для таких урезанных сигналов). Если после произнесения слова сигнал урезать таким же образом, то слова с меткой «-м» распознаются так же стопроцентно, как и исходные. Если оставить лишь урезанные эталоны, то слова будут стопроцентно распознаваться и без урезания распознаваемого сигнала.

Остановимся на используемой нами при DTW-распознавании системе признаков [2]. Отрезок речи, вводимый с микрофона, оцифровывается с частотой 22050 кГц. В соответствующий буфер заносится 10 тысяч чисел:

$$y_1, y_2, \dots, y_{10000} \tag{1}$$

значения напряжения на выходе микрофона в последовательные моменты времени (Эти моменты времени будем называть отсчетами). Сам ряд чисел (1) и соответствующую функцию

$$y(i) = y_i \tag{2}$$

будем называть сигналом. Таким образом, числа (1), в конечном счете, отражают изменение давления на мембрану микрофона как функцию времени. На экран монитора может быть выведен график сигнала, как функция времени (визуализация сигнала).

Сглаживанием сигнала мы называем обработку его 3-точечным скользящим фильтром

$$y_i = \frac{y_{i-1} + y_i + y_{i+1}}{3}, \quad i = 2, 3, \dots, 9999 \tag{3}$$

Дальнейшая работа происходит с поточечной разностью исходного и десятикратно сглаженного сигнала. Это позволяет в некоторой степени «очистить» его от индивидуального тембра говорящего и тем самым сделать шаг в направлении дикторонезависимости системы распознавания. Далее, если не оговорено противное, под сигналом будем понимать указанную разность и, чтобы не усложнять обозначений, считать, что (1) и (2) соответствуют именно ей.

Пусть l – число отсчетов между двумя соседними локальными максимумами функции (2) (назовем сужение функции на соответствующий интервал полным колебанием). Если максимумы не строгие, то под l будем понимать число отсчетов от начала первого максимума до начала второго. Определим величину z:

$$z = l \text{ при } 2 \le l < 20; \; z = 20 + \frac{l-20}{6} \text{ при } 20 \le l < 50; \; z = 25 + \frac{l-50}{10} \text{ при } 50 \le l < 90; \; z = 29 \text{ при } l \ge 90.$$

Ближайшее целое число, не превосходящее z, назовем длиной соответствующего полного колебания. Таким образом, длина полного колебания учитывается тем более точно, чем оно короче. Выделим участок сигнала и обозначим через n общее число полных колебаний на этом участке, через n_1 – число полных колебаний длины 2,..., через n_{28} – число полных колебаний длины 29.

Поставим в соответствие выделенному участку вектор

$$(x_1, \dots, x_{28}, \varepsilon), \tag{4}$$

где $x_k = n_k / n$, $k = 1, 2, \dots 28$, ε – отношение амплитуды (разность наибольшего и наименьшего значений) рассматриваемого участка сигнала к амплитуде всего сигнала. Величина ε вводится для того, чтобы надежно отделить паузу от значащей части сигнала, а нормировка ее делается, чтобы отвлечься от громкости произносимого.

Разобьем записанный сигнал в 10 тысяч отсчетов на отрезки по 368 отсчетов в каждом (удвоенный квазипериод основного тона для мужского голоса средней высоты). Для каждого из 27 полных отрезков вычислим вектор (4). Последний неполный отрезок просто отбросим. В результате мы представляем сигнал в виде траектории, то есть последовательности 27 точек в 29-мерном пространстве:

$$A = (a_1, a_2, \dots, a_{27}).$$

III. Далее мы условимся называть выделяемый нами участок межфонемного перехода дифоном. При этом отметим, что наш участок для каждого межфонемного перехода внутри слова имеет стандартную длину и короче того, что обычно понимается под дифоном – отрезок от середины предшествующего звука до середины следующего.

Перечислим теперь некоторые преимущества предлагаемого подхода.

1. При использовании дифонов появляется надежный способ различения между собой звуков [б], [г], [д].

Выдержка звонких взрывных согласных [б], [г], [д] включает два момента: во-первых, органы речи образуют полную смычку; во-вторых, напор воздуха ее прорывает. На рис. 2 – 4 приведены визуализации сигналов, соответствующих словам «САБО», «САГА» и «САДА» (родительный падеж слова «САД»), содержащих звонкие взрывные звуки [б], [г], [д].

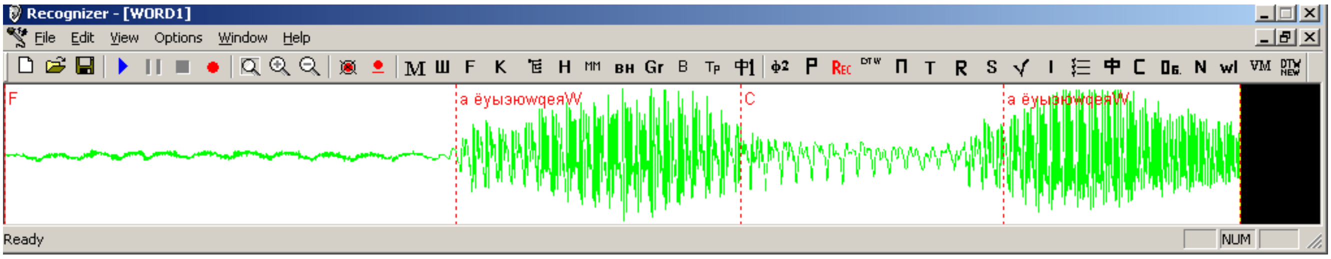


Рисунок 2 – Сигнал, отвечающий слову «сабо»

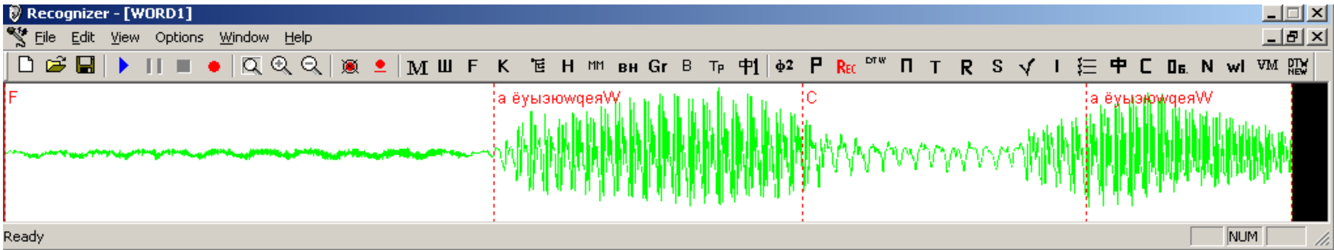


Рисунок 3 – Сигнал, отвечающий слову «сага»

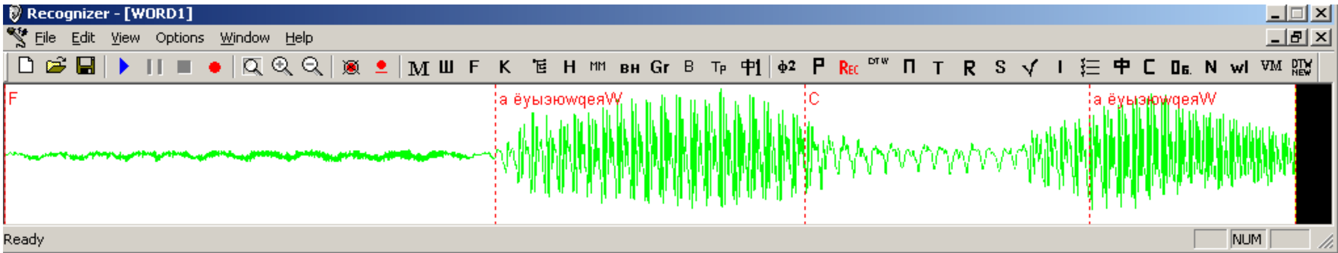


Рисунок 4 – Сигнал, отвечающий слову «сада»

Из них видно, что большая часть участков, отвечающих этим звукам, является квазипериодической. Этот факт легко понять, если попытаться произнести эти звуки изолированно, без последующего гласного. Мы видим, что еще до взрыва начинают звучать голосовые связки, они и создают квазипериодический отрезок в сигнале. Отличия же между указанными звуками в приведенных словах сосредоточены на очень коротком переходе к последующему звуку. Поэтому, работая со стационарными частями, мы были лишены возможности различать указанные звуки между собой. При использовании межфонемных переходов такая возможность появляется.

2. Появляется надежный способ различения между собой звуков [к], [п], [т] в середине слова. Пример – слова «папа» и «пата» (родительный падеж от шахматного термина «пат»). На рис. 5 приведена визуализация сигнала, отвечающего слову «ЛАПА».

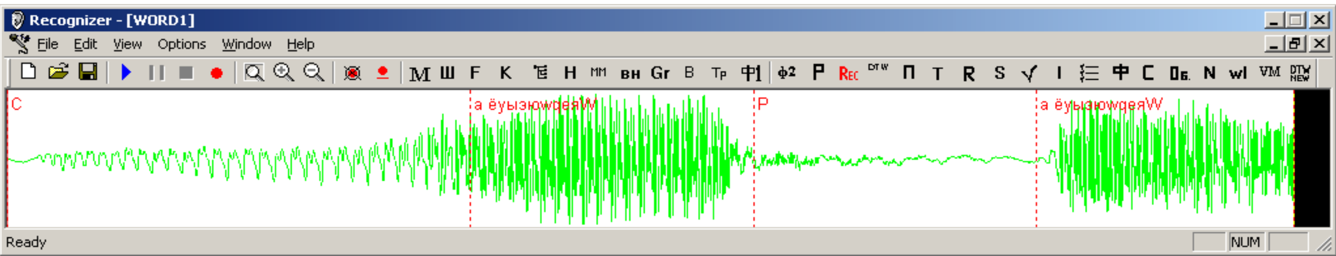


Рисунок 5 – Сигнал, отвечающий слову «лапа»

Поскольку при произнесении глухих взрывных [к], [п], [т] также есть момент полного перекрытия голосового тракта, но голосовые связки в этот момент молчат, то в сигнале появляется характерный паузообразный сегмент. Работая со стационарными частями, мы были лишены возможности различать между собой и эти звуки. С использованием дифонов такая возможность появляется.

3. Становятся надежно различимыми парные твердые и мягкие согласные. К связанному с этим вопросу вернемся в конце пункта IV.

4. Появляются новые возможности в распознавании сверхбольших словарей. На последнем остановимся подробнее.

IV. Имея достаточно совершенную систему автоматической сегментации (разбиения записанного слова на участки, отвечающие отдельным звукам) и автоматического отнесения каждого звука к гласным (у начальной метки участка проставляется идентификатор W), звонким согласным (C), шипящим (F) и паузообразным звукам (P), мы разработали систему, которая, используя стационарные части звуков, классифицировала их в рамках указанных классов. Она должна была различать между собой гласные, различать между собой звонкие согласные и различать между собой шипящие и свистящие звуки. В ходе этого весьма частыми были отказы от распознавания, что приводило к необходимости вместо определения одного из двух звуков допускать возможность присутствия каждого из них. (Напомним, что мы с самого начала отказывались от распознавания между собой [б], [г], [д], а также от различения между собой [к], [п], [т]). В результате вместо конкретного распознанного слова мы получали список слов – кандидатов на распознавание. Отметим, что его размер на порядки меньше размера исходного словаря. В полученном списке пользователь двойным щелчком мыши выделял нужное слово. При этом автоматически создавался голосовой эталон с именем соответствующей *леммы* (словарная форма слова), который позволял в последующем, используя DTW-распознавание в пределах списка кандидатов, в большинстве случаев распознавать словоформы этого слова, отождествляя их с данной леммой. В случае, когда упомянутый список кандидатов сводился к одному слову, эталон для него создавался без дополнительного указания пользователя. Таким образом, исключая эти не часто встречавшиеся случаи, пользователю приходилось, обучая программу по ходу распознавания, самому создавать эталоны большинства произносимых слов.

При новом подходе достаточно создать эталоны всех дифонов, что при наличии удобной программы пользователь может сделать за один сеанс работы (такая программа описана ниже). Это оказывается достаточным ввиду следующего.

У нас есть программа, которая по написанному русскому слову создает его транскрипцию и синтезирует эталон этого слова из эталонов дифонов. На рис. 6 показан фрагмент окна, связанного с этими функциями. По нажатии кнопки «синтез» при выключенном флажке «синтез из эталонов» из звуковых файлов дифонов «склеивается» звуковой файл слова и создается его DTW-эталон. При включенном флажке эталон слова склеивается непосредственно из эталонов дифонов. Последний вариант является более быстрым и обладает тем преимуществом, что составляющие эталоны дифонов можно усреднять [2]. Последнее важно при работе в направлении дикторонезависимости.

Возвратимся к программе распознавания большого словаря. Теперь мы получаем возможность, пользуясь сравнительно небольшой величиной списка слов – кандидатов на распознавание, автоматически синтезировать эталоны этих слов из эталонов дифонов и вести DTW-распознавание по этим эталонам в пределах указанного списка. Результат – однозначное распознавание слова без дополнительного вмешательства пользователя.

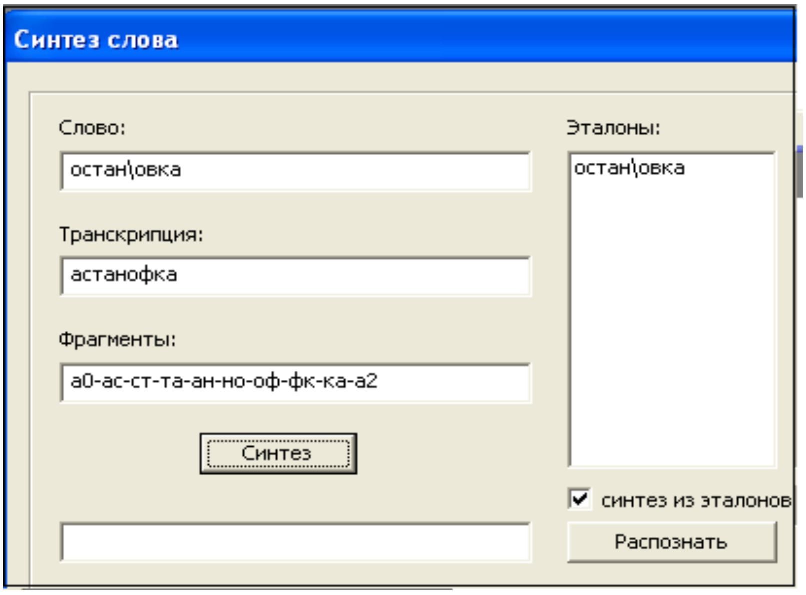


Рисунок 6 – Фрагмент окна для синтеза.
Синтезирован из дифонов эталон слова «остановка»

Отметим, что при распознавании слов словаря Зализняка [4], содержащего около 94600 слов, достаточно работать со списками кандидатов на распознавание, которые получаются при сегментации и широкой фонетической классификации (классы W, C, F, P), если сократить этот список путем классификации дифонов типа CW, FW, PW по твердости-мягкости. Остановимся на последней процедуре подробнее. Для определенности будем говорить о дифонах типа CW.

Для классификации по твердости-мягкости используется перечень дифонов [ба], [ва]…[bи], [vи],…, где латинскими буквами обозначены мягкие звонкие согласные звуки. Здесь довольно много объектов, но при распознавании учитываются только два класса: твердый дифон и мягкий дифон. То есть результаты [ба], [ва], [га],…,[но], [ну], [ны], [нэ] отождествляются. Отождествляются также результаты [bё],…,[nq] (q- транскрипционный символ для ударного звука [я]). Аналогично анализируются на твердость – мягкость дифоны типа FW, PW. Такой способ распознавания на «твердость – мягкость», как показывает опыт, обладает очень высокой надежностью.

При описанной работе со словарем Зализняка слово «мама», например, попадает в список из 179 слов со структурой CWCW, для которых фактически в реальном времени вышеописанным образом создаются надежные эталоны.

Таким образом, при новом подходе к распознаванию больших словарей ситуация меняется радикально. Вместо того, чтобы, в конечном счете, создавать эталоны для всех слов распознаваемого словаря, пользователю достаточно создать эталоны для дифонов (межфонемные переходы). При этом распознаваемые словари могут насчитывать сотни тысяч и даже миллионы словоформ русских слов, в то время как количество всех дифонов по порядку величины есть квадрат от количества транскрипций звуков русской речи, числом всего около 1450.

Сделаем дополнительные замечания. Спрашивается, зачем синтезировать из эталонов дифонов эталон слова? Не проще ли, распознавая дифоны между собой, получить их список, соответствующий слову, и по нему распознать слово? Нет, не проще. Дело в том, что дифонов достаточно мало для обучения, но слишком много для распознавания словаря этих дифонов, в особенности учитывая, что это короткие звуковые единицы, а DTW-распознавание, естественно, тем надежнее, чем длиннее и разнообразнее по составу распознаваемые объекты. Создание синтезированных эталонов слов – как раз шаг в этом направлении и он позволяет использовать DTW-распознавание целых слов со всеми его преимуществами.

V. Остановимся на программе, которая позволяет достаточно быстро создать эталоны всех дифонов.

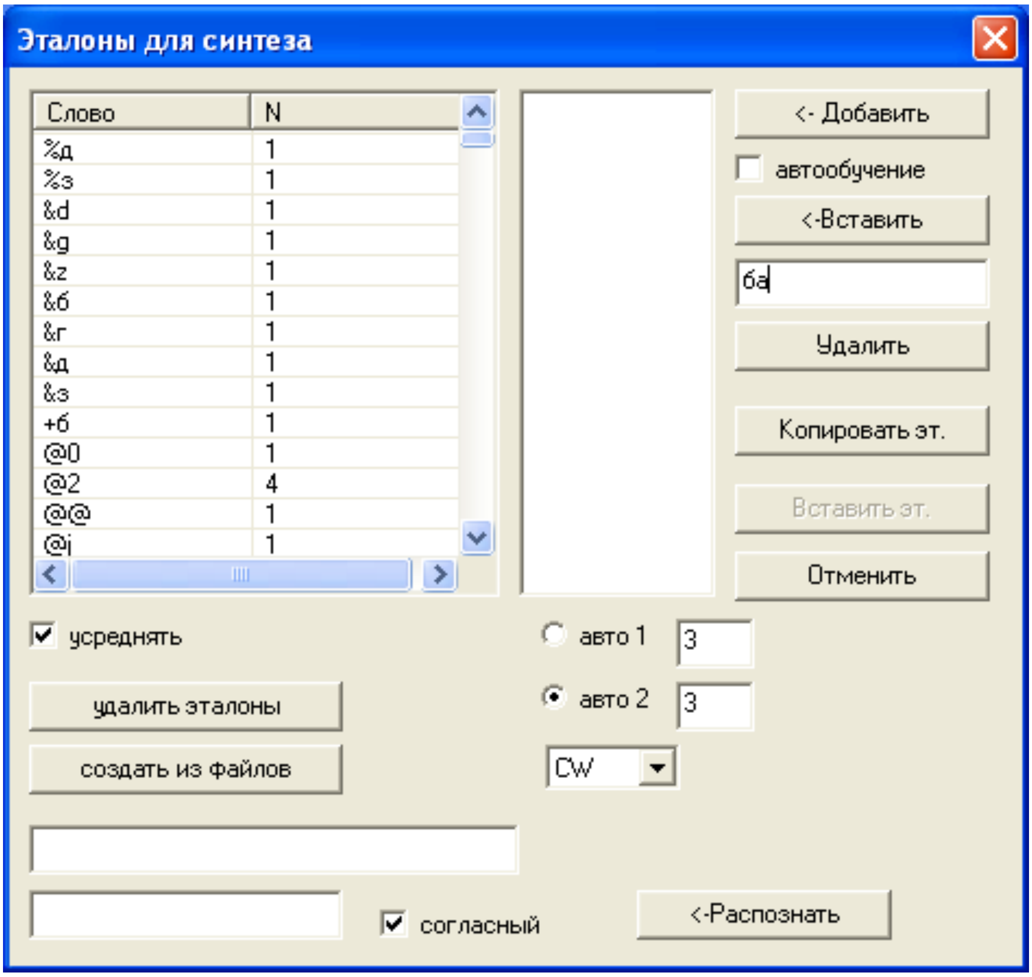


Рисунок 7 – Окно для создания базы дифонов

На рис. 8 показана визуализация сигнала, отвечающего слову «собака» с автоматически выделенным дифоном «ба». Рис. 7 – окно программы для создания базы дифонов. Программа работает следующим образом. В поле, в котором на рисунке записано «ба», записывается имя нужного дифона или участка звука в начале или в конце слова (имена последних снабжаются в конце символами 0 или 2 соответственно). При этом происходит автоматическое включение радио-кнопки «авто1» (выделяется участок в начале слова) или «авто2» (выделяется дифон в середине слова), и в выпадающем списке

WC, WF, WP, CW, CC, CF, CP, FW, FC, FF, FP, PW, PC, PF, PP,

(на рис. 7 в соответствующем окне записано CW) автоматически выбирается тип межфонемного перехода.

При произнесении слова сигнал автоматически сегментируется. В окне, представленном на рис. 8, проставляются метки, отделяющие участки соседних звуков.

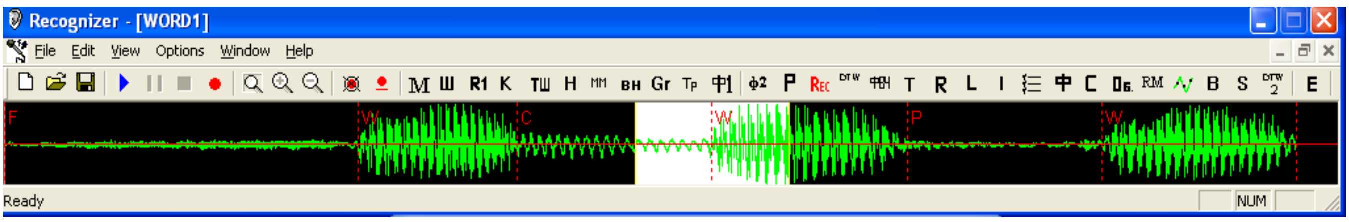


Рисунок 8 – Сигнал, отвечающий слову «собака»
с автоматически выделенным дифоном «ба»

Далее автоматически выделяется окрестность метки в первом дифоне слова, имеющем указанный тип, и по нажатии кнопки «Вставить» создается его эталон, который помещается в базу эталонов. Участки в конце слова пока выделяются руками с помощью мышки.

Если записать в упомянутом поле имя дифона, затем включить флажок «автообучение» и проделать вышеописанные операции, то после нажатия кнопки «вставить» в этом поле появится имя следующего дифона из заранее подготовленного списка дифонов и будут выполнены операции, позволяющие автоматически выделить его в сигнале и создать его эталон после произнесения подходящего звукосочетания. Именно эти возможности позволяют пользователю создать базу эталонов дифонов, что называется, «в один присест».

Литература

1. О распознавании фонем с помощью анализа речевого сигнала в частотной и временной областях. Приложение к распознаванию синтаксически связных фраз / В.Ю. Шелепов, А.В. Ниценко, А.В. Жук [и др.] // Речевые технологии. – 2008. – № 2. – С. 43-52.
2. Шелепов В.Ю. Лекции о распознавании речи / Шелепов В.Ю. – Донецк : ІПШІ «Наука і освіта», 2009. – 196 с.
3. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов / Винцюк Т.К. – Киев : Наукова думка, 1987. – 262 с.
4. Зализняк А.А. Грамматический словарь русского языка / Зализняк А.А. – М. : Русский язык, 1977. – 879 с.

Literatura

1. Shelepov V.Ju., Nicenko A.V., Zhuk A.V.. Rechevye tehnologii. 2008. № 2. S. 43-52.
2. Shelepov V.Ju. Lekcii o raspoznavanii rechi. Doneck: IPShI "Nauka i osvita". 2009. 196 s.
3. Vincjuk T.K. Analiz, raspoznavanie i interpretacija rechevyh signalov. Kiev: Naukova dumka, 1987. 262 s.
4. Zaliznjak A.A. Grammaticheskij slovar' russkogo jazyka. M.: Russkij jazyk, 1977. 879 s.

V.Ju. Shelepov, A.V. Nicenko, G.V. Dorohina

On Speech Recognition Using Phoneme Transitions Base

We can begin the analysis of this statement, with the next simple experiment. Using any known program for work with sounds, "Sound Forge" for example, let's record two random words and cut stationary (central) phoneme parts. After reproduction we can recognize the phonation of these words. On the contrary, if we cut phoneme transitions and keep stationary parts, we will be at loss for recognition of such words as "mama" and "lama".

The next argument is use of samples after cut all phoneme stationary parts for DTW-recognition ([1], see also [2]). The experiments shows, that such recognition is not less successful then recognition of full samples.

Let's name a separated part of phoneme transition "diphone". Our diphone has a standard length and is shorter then usual diphone, which is a piece from the middle phoneme area to the middle of the next phoneme area.

Here is the list of advantages of the proposed approach:

1) When using diphones we have an opportunity for recognition of phonemes [b], [g], [d] and recognition of phonemes [p], [k], [t].

2) Pairs of hard and soft consonants in the Russian language become distinguishable.

3) New opportunities for recognitions of super large vocabularies are appearing. Let's describe this advantage in more detail.

When using stationary phoneme parts, recognition breakdowns in phoneme recognition when using stationary phoneme parts often take place. Because of this fact, instead of determination of one or two phonemes, it is necessary to entertain a possibility that each of the phonemes is present. As the result, instead of recognized word, we have a list of candidates for recognition. Its size is one order less then initial vocabulary size.

A user should choose a proper word in the list by double-click. At the same time, voice sample with the name of the main form of word was created. It permitted later (in most cases) to recognize all forms of the word identifying them with main form. When the list of candidates contains only one word, the sample for it was created automatically.

Thus a user had to create the samples for the most part of the pronounced words.

In the new approach, it is sufficient to create samples for all diphones. We have a comfortable program, which allows doing it at once. It is enough because of the following facts.

We have a program, which is written for Russian word and which forms the transcription for Russian words and syntheses their sample using diphone samples. So we get the opportunity (in view of not big size of list of candidates for recognition) to synthesize samples for these word candidates automatically, using diphone samples and doing DTW-recognition within this list.

The result is unique recognition of the word without additional user's intervention.

Thus in the new recognition approach, the situation radically changes. Instead of creating samples for all recognized words, it is sufficient to create samples only for diphones. At the same time, vocabulary can contain thousands and even millions of forms of Russian words while quantity of all diphones according to the order of value is the second degree of quantity Russian phonemes (the total sum is about 1450).

We also must note that for recognition of all words of Zaliznjak's dictionary, which contains about 94600 words, it is sufficient to work only with wide phonetic classification if to reduce the lists of candidates using hard and soft diphone classification. It is done with a high level of reliability.

Статья поступила в редакцию 01.04.2011.